Supplementary information

Somatic deletions implicated in functional diversity of brain cells of individuals with schizophrenia and unaffected controls

Junho Kim[1], Jong-Yeon Shin [2, 3], Jong-Il Kim [2,3,4,5], Jeong-Sun Seo [2,3,4,5,6] , Maree J. Webster [7], Doheon Lee[1]* and Sanghyeon Kim[7]*

[1]Department of Bio and Brain Engineering, KAIST**,** 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Korea

[2]Genomic Medicine Institute (GMI), Medical Research Center, Seoul National University, Seoul 110-799, Korea

[3]Psoma Therapeutics Inc., Seoul, 153-781, Korea

[4]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul 110-799, Korea

[5]Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Seoul 110-799, Korea

[6]Macrogen Inc., Seoul 153-781, Korea

[7]Stanley Brain Research Laboratory, Stanley Medical Research Institute, 9800 Medical Center Drive, Rockville, MD 20850

*Corresponding authors

Sanghyeon Kim, PhD
Stanley Medical Research Institute
9800 Medical Center Drive
Rockville, MD 20850
Email: kims@stanleyresearch.org

Doheon Lee, PhD
KAIST
291 Daehak-ro
Yuseong-gu, Daejeon 305-701
Email: dhlee@kaist.ac.kr

Supplementary Table S1. Demographic and clinical information for the cases and controls

ID[1]	Profile	Age	Sex	PMI	Brain PH	Alcohol	Drug	Smoking	Psychotic Feature	Lifetime Antipsychotics
A9	schizophrenia	45	F	52	6.51	4	4	unknown	yes	20000
C16	schizophrenia	60	M	31	6.2	0	0	unknown	yes	80000
C17	schizophrenia	32	M	19	6.1	5	5	yes	yes	15000
C13	control	52	M	28	6.5	2	0	No	No	0
C21	control	59	M	26	6.4	3	0	yes	No	0

A, Array collection; C, Stanley neuropathology consortium; PMI, postmortem interval, Scale for alcohol and drug use, 0=Little/None, 1=social, 2=Moderate in past, 3=Moderate in present, 4=Heavy in past, 5=Heavy in present.

Supplementary Table S2. Somatic deletion candidates in DNA from brain of an individual with schizophrenia (A9) that were called using the pipeline for data from multiple tissues.

Tissue	Chr	Start	Size (bp)	Gene1	Gene2
PFC	chr2	90961149	755		
PFC	chr2	179023695	500		*MIR548N,PRKRA*
PFC	chr2	120133653	1241		
PFC	chr7	145299659	74170		
PFC	chr9	134748910	997		C9orf9
PFC	chr14	39057228	4711		
PFC	chr19	14824186	51155	OR7A17	
C	chr7	26212701	1818		CBX3
C	chr18	98677	440		
C	chr21	21705368	816		NCAM2
Common	chr3	196024570	2926		
Common	chr5	172967733	1304		*BOD1*
Common	chr6	119119028	1646		CEP85L
Common	chr6	136631214	1052		BCLAF1
Common	chr7	26214778	3290		*CBX3*
Common	chr10	41678417	5612		
Common	chr11	128518741	1786		ARHGAP32
Common	chr21	9755944	538		
Common	chr21	9771916	2258		
Common	chr21	13287780	746		

Somatic deletions in brain DNA; chromosomal annotation (hg18); Gene1; deleted genes; Gene2; genes disrupted by breakpoints; PFC, frontal cortex; C, cerebellum; Common, both brain regions

Supplementary Table S3. Somatic deletion candidates in DNA from blood of an individual with schizophrenia (A9) that were called using the pipeline for data from multiple tissues.

Tissue	Chr	Start	Size (bp)	Gene*
Blood	chr6	32620415	466	
Blood	chr6	136641508	3788	*BCLAF1*
Blood	chr7	156467913	579	
Blood	chr8	98897657	35112	*LAPTM4B*
Blood	chr11	48308185	562	
Blood	chr11	133375270	516	
Blood	chr12	102902805	1950	*TDG*
Blood	chr12	102903637	1342	*TDG*
Blood	chr15	18368022	506	
Blood	chr17	219937	455	
Blood	chr19	23441261	595	
Blood	chr22	34419487	421	

Chromosomal annotation (hg18); Gene*; genes disrupted by deletions.

Supplementary Table S4. Summary of validation results in the first discovery phase and the second phase

	First phase		Second phase
	Before BLAT and size filter	After BLAT and size filter	
Total validated deletions	7(3)[1]		12(3)[1]
Total False positives	12(8)[2]	3(2)[2]	1(1)[2]

[1] The numbers in parenthesis show the number of additionally confirmed somatic deletions with different breakpoints in different tissue or cells from the originally called somatic candidates. [2]The numbers in parenthesis show the number of false positives which were miscalled due to the variations in the calling pipelines.

Supplementary Table S5. Somatic deletion candidates in the PFC DNA of an individual with schizophrenia A9 that were called using the pipeline for data from a single tissue

Tissue	Chr	Start	End	Size (bp)	Gene1	Gene2
PFC*	chr2	90961149	90961279	755		
PFC*	chr2	120133653	120134976	1241		
PFC*	chr2	179023362	179023998	560		*MIR548N,PRKRA*
PFC*	chr7	145299659	145373837	74170		
PFC*	chr9	134748910	134749961	997		*C9orf9*
PFC*	chr14	39057228	39062055	4711		
PFC*	chr19	14824186	14875314	51155	*OR7A17*	
PFC	chr2	109182231	109182767	511		*SH3RF3*
PFC	chr2	132700574	132704545	4077		*ANKRD30BL*
PFC*	chr3	196024570	196027580	2926		
PFC	chr4	48815807	48816253	464		
PFC*	chr6	136631214	136632284	1052		*BCLAF1*
PFC*	chr7	26214746	26218225	3476		*CBX3*
PFC	chr10	27264250	27268097	3866		*LINC00202*
PFC	chr10	38856514	38858168	1535		
PFC	chr12	92418926	92419817	885		*MRPL42*

Chromosomal annotation (hg18); Gene1; deleted genes; Gene2; genes disrupted by breakpoints; PFC, frontal cortex; PFC*, candidates were called somatic deletions specific to PFC or common to PFC and cerebellum when we used the pipeline for multiple tissue data.

Supplementary Table S6. Somatic deletions in prefrontal cortex DNA of two schizophrenic samples and two unaffected control samples

Sample ID	Chr	Start	End	Size (bp)	Gene
C13	chr1	16882167	16885398	3215	
C13	chr2	179009533	179016237	6865	MIR548N,PRKRA
C13	chr3	138899570	143942728	5043127	A4GNT,ACPL2,ARMC8,ATP1B3,ATR,BPESC1,C3orf72,CEP70,CLDN18,CLSTN2,COPB2,DBR1,DZIP1L,ESYT3,FAIM,FOXL2,GK5,GRK7,MRAS,MRPS22,NME9,NMNAT3,PIK3CB,PISRT1,PLS1,PRR23A,PRR23B,PRR23C,RASA2,RBP1,RBP2,RNF7,SLC25A36,SOX14,SPSB4,TFDP2,TRIM42,TRPC1,XRN1,ZBTB38
C13	chr4	31709834	31714948	5086	
C13	chr4	48789879	48848899	58987	
C13	chr4	105276717	105464303	187584	
C13	chr4	179491177	179491888	921	
C13	chr5	98895914	98896989	996	
C13	chr5	115205763	115233612	27840	AP3S1
C13	chr6	82201214	82235796	34674	
C13	chr6	136641919	136642449	993	BCLAF1
C13	chr7	139828224	139836872	8579	
C13	chr8	12472468	12477236	5034	
C13	chr9	69242893	69243387	508	
C13	chr10	66531974	66954519	422471	
C13	chr11	48324153	48330194	6192	OR4C45
C13	chr11	50674033	50676261	2376	
C13	chr12	33897305	34525668	628292	ALG10
C13	chr12	102897955	102900722	2682	TDG
C13	chr13	107707087	108247256	540130	MYO16,TNFSF13B
C13	chr14	36701434	36841011	139635	LOC100129794,MIPOL1,SLC25A21
C13	chr14	105589267	106253456	664993	LINC00221,LINC00226
C13	chr15	39640893	39642002	1169	TYRO3
C13	chr15	39652090	39652798	688	TYRO3
C13	chr16	33892902	33893672	1175	

C13	chr17	25436676	27070069	1633505	ADAP2,ATAD5,BLMH,CPD,CRLF3,DPRXP4,EFCAB5,EVI2A,EVI2B,GOSR1,LRRC37BP1,MIR193A,MIR3184,MIR365B,MIR423,MIR4724,MIR4725,MIR4733,NF1,NSRP1,OMG,RAB11FIP4,RNF135,SH3GL1P2,SLC6A4,SUZ12P,TBC1D29,TEFM,TMIGD1
C13	chr19	980045	980495	430	CNN2
C13	chr19	57878459	58173711	295215	ZNF28,ZNF320,ZNF321P,ZNF468,ZNF600,ZNF611,ZNF702P,ZNF816,ZNF816-ZNF321P,ZNF83
C13	chr19	59185658	59186150	479	
C21	chr1	91334093	93287175	1953179	BRDT,BTBD8,C1orf146,CDC7,EPHX4,EVI5,FAM69A,GFI1,GLMN,HFM1,HSP90B3P,KIAA1107,RPAP2,RPL5,SNORA66,SNORD21,TGFBR3
C21	chr1	142320289	142365119	44783	
C21	chr1	143612541	143618200	5631	PDE4DIP
C21	chr2	70514828	70515957	1076	
C21	chr4	48798450	48799175	778	
C21	chr4	48789084	48802152	13145	
C21	chr4	48789873	48848898	58989	
C21	chr6	136632444	136635802	3404	BCLAF1
C21	chr7	26214791	26217983	3465	CBX3
C21	chr7	102587817	102588581	715	
C21	chr8	131803905	142724228	10920268	ADCY8,CHRAC1,COL22A1,DENND3,EFR3A,EIF2C2,FAM135B,FLJ43860,GPR20,HHLA1,HPYR1,KCNK9,KCNQ3,KHDRBS3,LOC286094,LOC731779,LRRC6,MIR30B,MIR30D,NDRG1,OC90,PHF20L1,PTK2,PTP4A3,SLA,SLC45A4,ST3GAL1,TG,TMEM71,TRAPPC9,WISP1,ZFAT,ZFAT-AS1
C21	chr9	66272182	66273128	915	
C21	chr10	5192150	5314501	122655	AKR1C4,AKR1CL1
C21	chr12	33897310	34525592	628318	ALG10
C21	chr12	102902818	102903289	679	TDG
C21	chr14	36701355	36840977	139614	LOC100129794,MIPOL1,SLC25A21
C21	chr16	8554261	8554889	755	
C21	chr19	21399402	21400252	824	ZNF493

C16	chr1	121053571	121180300	126667	
C16	chr4	75984	356949	281189	HMX1,MGC39584,ZNF141,ZNF595,ZNF718,ZNF732,ZNF876P
C16	chr4	48804428	48810991	6845	
C16	chr4	118808950	118822481	13496	
C16	chr7	36383216	36563156	179902	ANLN,AOAH,KIAA0895
C16	chr7	52138621	52364875	226267	
C16	chr9	69242776	69243373	532	
C16	chr11	93607641	93613403	5664	
C16	chr12	33897348	34525567	628313	ALG10
C16	chr15	39642298	39644383	2242	TYRO3
C16	chr15	39647022	39647705	651	TYRO3
C16	chr15	39652214	39652796	662	TYRO3
C16	chr22	15366199	15367015	836	
C16	chrX	55189439	55202175	12842	FAM104B
C16	chr2	179023593	179023955	546	MIR548N,PRKRA
C17	chr1	89371862	89424739	53049	GBP4,GBP7
C17	chr1	121087982	121181573	93667	
C17	chr2	70514820	70515956	1062	
C17	chr3	67576372	67579541	3113	SUCLG2
C17	chr4	48789874	48848829	58999	
C17	chr4	48817156	48838173	20966	
C17	chr5	1716735	1717276	479	
C17	chr7	6356161	6360883	4991	
C17	chr7	52138602	52364750	226245	
C17	chr9	21971233	25396804	3425692	CDKN2A,CDKN2B,CDKN2B-AS1,DMRTA1,ELAVL2,FLJ35282
C17	chr9	66272267	66273086	893	

C17	chr11	116194546	124400193	8205749	ABCG4,AMICA1,APOA1,APOA4,APOC3,ARCN1,ARHGEF12,ATP5L,BACE1,BACE1-AS,BCL9L,BLID,BSX,C11orf61,C11orf63,C1QTNF5,C2CD2L,CBL,CCDC15,CCDC153,CCDC84,CD3D,CD3E,CD3G,CEP164,CLMP,CRTAM,CXCR5,DDX6,DPAGT1,DSCAML1,ESAM,FOXR1,FXYD2,FXYD6,FXYD6-FXYD2,GRAMD1B,GRIK4,H2AFX,HEPACAM,HEPN1,HINFP,HMBS,HSPA8,HYOU1,IFT46,IL10RA,LOC100499227,LOC100526771,LOC100652768,LOC341056,LOC649133,MCAM,MFRP,MIR100,MIR100HG,MIR125B1,MIR3656,MIR4492,MIR4493,MIRLET7A2,MLL,MPZL2,MPZL3,NLRX1,NRGN,OAF,OR10G4,OR10G7,OR10G8,OR10G9,OR10S1,OR4D5,OR6M1,OR6T1,OR6X1,OR8A1,OR8B12,OR8B2,OR8B3,OR8B4,OR8B8,OR8D1,OR8D2,OR8D4,OR8G1,OR8G2,OR8G5,PAFAH1B2,PANX3,PCSK7,PDZD3,PHLDB1,POU2F3,PVRL1,RNF214,RNF26,ROBO3,ROBO4,RPL23AP64,RPS25,SC5DL,SCN2B,SCN3B,SCN4B,SIAE,SIDT2,SIK3,SLC37A4,SORL1,SPA17,TAGLN,TBCEL,TBRG1,TECTA,THY1,TMEM136,TMEM225,TMEM25,TMPRSS13,TMPRSS4,TRAPPC4,TREH,TRIM29,TTC36,UBASH3B,UBE4A,UPK2,USP2,VPS11,VSIG2,VWA5A,ZNF202
C17	chr12	102903625	102904795	1224	TDG
C17	chr13	107707162	108247285	540102	MYO16,TNFSF13B
C17	chr16	44944405	44957443	13532	
C17	chr18	63189198	63749778	560541	DSEL,LOC643542
C17	chr19	21399375	21400139	832	ZNF493
C17	chrX	52904496	52906755	2322	

chromosomal annotation (hg18); Gene; deleted genes or genes disrupted by breakpoints

Supplementary Table S7. Biological processes over-represented in genes disrupted by somatic deletions in prefrontal cortex DNA of two schizophrenia cases.

Biological process	Count	Genes	PValue	FDR
GO:0007608~sensory perception of smell	20	OR10S1, OR8G2, OR8G5, OR6T1, OR10G4, OR4D5, OR6M1, OR10G7, OR10G8, OR10G9, OR8B8, OR8B12, OR8B2, OR8B3, OR8D1, OR8B4, OR8A1, OR8D2, OR8G1, OR8D4, OR6X1	1.64E-09	1.70E-06
GO:0007606~sensory perception of chemical stimulus	20	OR10S1, OR8G2, OR8G5, OR6T1, OR10G4, OR4D5, OR6M1, OR10G7, OR10G8, OR10G9, OR8B8, OR8B12, OR8B2, OR8B3, OR8D1, OR8B4, OR8A1, OR8D2, OR8G1, OR8D4, OR6X1	8.98E-09	4.65E-06
GO:0007600~sensory perception	23	TECTA, OR10S1, OR8G2, OR8G5, OR6T1, MFRP, OR10G4, OR4D5, OR6M1, C1QTNF5, OR10G7, OR10G8, OR10G9, OR8B8, OR8B12, OR8B2, OR8B3, OR8D1, OR8B4, OR8A1, OR8D2, OR8D4, OR8G1, OR6X1	4.73E-07	1.63E-04
GO:0050890~cognition	23	TECTA, OR10S1, OR8G2, OR8G5, OR6T1, MFRP, OR10G4, OR4D5, OR6M1, C1QTNF5, OR10G7, OR10G8, OR10G9, OR8B8, OR8B12, OR8B2, OR8B3, OR8D1, OR8B4, OR8A1, OR8D2, OR8D4, OR8G1, OR6X1	3.23E-06	8.35E-04

GO:0007186~G-protein coupled receptor protein signaling pathway	25	OR10G4, OR10G7, OR10G8, APOA1, CXCR5, OR10G9, APOC3, OR8A1, OR8G1, OR10S1, CD3E, OR8G2, OR8G5, OR6T1, ARHGEF12, OR4D5, OR6M1, OR8B8, OR8B12, OR8B2, OR8B3, OR8D1, OR8D2, OR8B4, OR8D4, OR6X1	9.10E-06	0.002
GO:0050877~neurological system process	26	GRIK4, OR10G4, MFRP, C1QTNF5, OR10G7, OR10G8, PVRL1, OR10G9, OR8A1, OR8G1, TECTA, SCN2B, OR10S1, OR8G2, OR8G5, OR6T1, OR4D5, OR6M1, OR8B8, OR8B12, OR8B2, OR8B3, OR8D1, OR8D2, OR8B4, OR8D4, OR6X1	1.02E-05	0.002
GO:0007166~cell surface receptor linked signal transduction	31	GRIK4, OR10G4, OR10G7, OR10G8, APOA1, CXCR5, OR10G9, APOC3, OR8A1, OR8G1, CD3G, CD3D, OR10S1, CD3E, OR8G2, OR8G5, CBL, OR6T1, ARHGEF12, THY1, OR4D5, OR6M1, OR8B8, OR8B12, OR8B2, OR8B3, OR8D1, OR8B4, OR8D2, PDZD3, OR8D4, OR6X1	1.28E-04	0.02

Supplementary Table S8. Mean insert sizes, standard deviations of the insert sizes and minimal size of detectable deletions

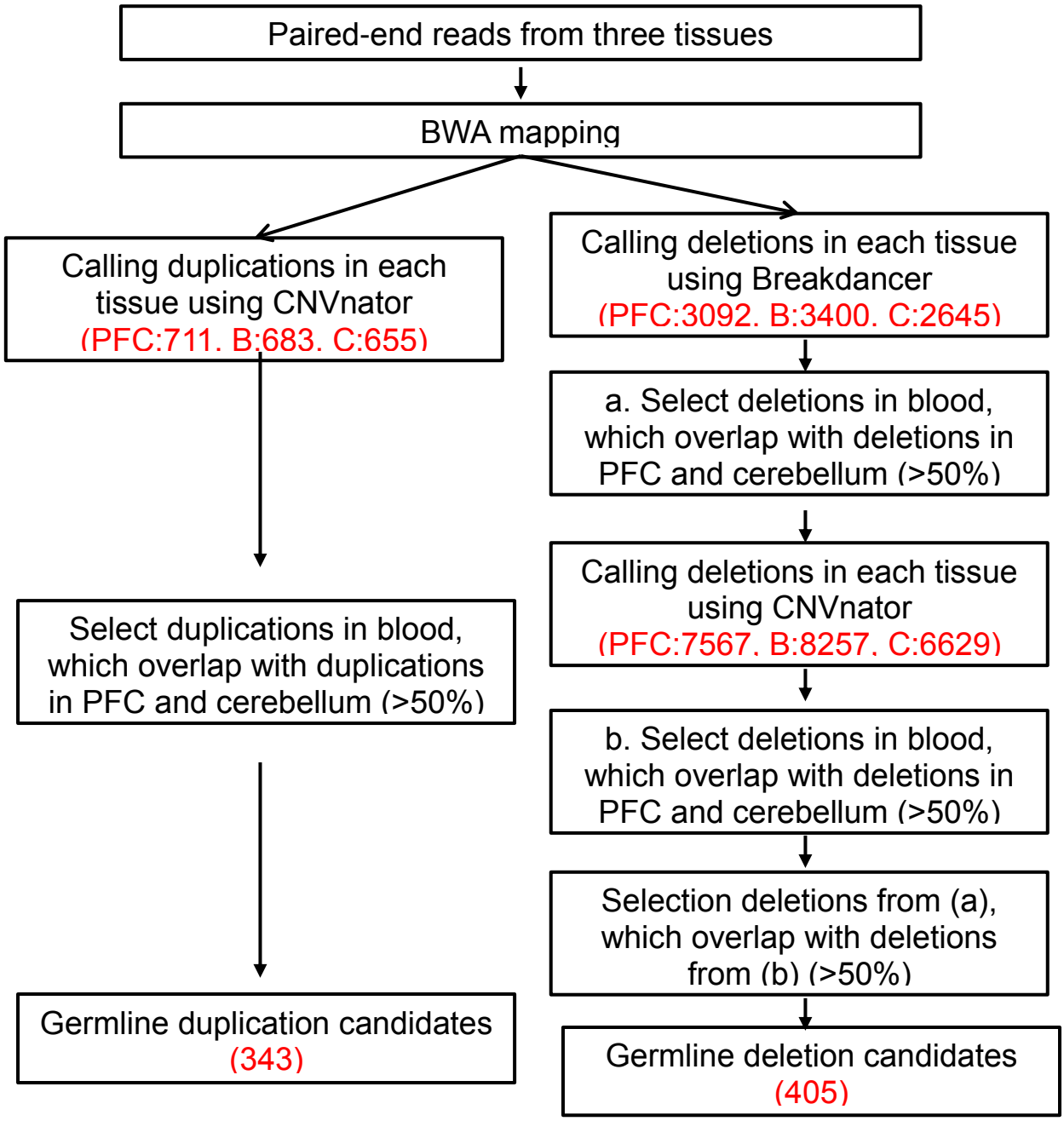
ID	Tissue	Mean	Standard deviation	Minimal size of detectable deletions (bp)
A9	PFC	330.4	79.86	258
A9	Cerebellum	319.8	85.54	283
A9	Blood	360.5	97.41	285
C13	PFC	302.9	43.53	128
C16	PFC	306.2	57.74	144
C17	PFC	310.6	49.7	129
C21	PFC	281.4	47.29	112

SupplementaryTable S9. Primers used for validating germline deletions and somatic CNVs.

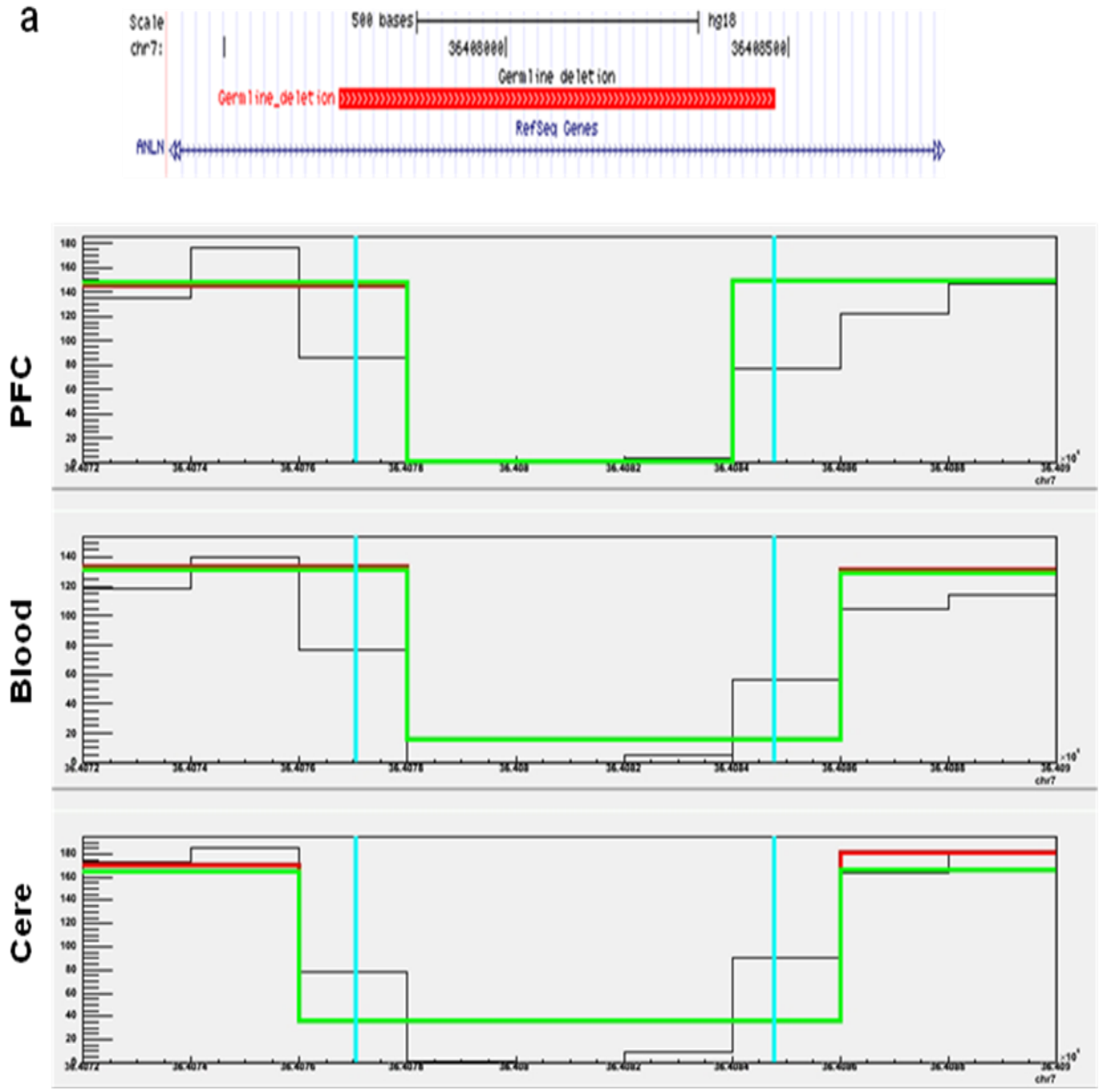
Type	Location (hg18)	Gene	Primer F	Primer R	Primer nested F	Primer nested R
Germline deletion	chr4: 6472894-6474412	PPP2R2C	GACAGTCTGCAATCCCCAAT	CCTCATCAACCTCCCAAGTC		
Germline deletion	chr7:36407703-36408477	ANLN	ACTCCTGTGACCTGCCTGAT	TGGAGGAGGGAGAGACTCCT		
Germline deletion	chr14:64625312 - 64627072	MAX	ATTGCCCAAGTTGGAGTCTG	TGCGTCCTAAGCCTTTTGTT		
Germline deletion	chr15: 97171281-97183365	IGF1R	GGAGCAATGCTGAACCACTT	CACCAAAAGCTGGAGCAACT		
somatic control	chr3:58278773-58278954	RPP14	AAAGCTGAAGCGGTTCATTG	AGCAATTCCCCATAGGCTCT		
Somatic duplication	chr1:150589801-150596600	FLG2	ACTTGTGGTTGGACCTGAGC	GGCTTTGCACAGCATGAGTA		

Somatic duplication	chr15:19461601-19466200	LOC348120	CCCAATACATGTGTGGCAAA	GGGGCTACTTCTAAACTTCTATCA		
Somatic duplication	chr1:150296801-150302600	intergenic	TGGGATTTGGTTTCTGCTCT	TCCCAACATAGGGTCCAGAA		
Somatic duplication	chr19:20840401-20891200	intergenic	AGCCATGCAGTCACCACATT	CAGGTTCTCAGGGTGTTGAA		
Somatic deletion	chr19:10031001-10034000	C3P1	GACACAAAGAGAGGGCAGGT	GCGTGTGTCTCATTGTACCG		
Somatic deletion	chr7:150399401-150401800	SLC4A2_L1	TCTGTAGCAGCAACCACCTG	CGAAGATCTCCTGGGTGAAG		
Somatic deletion	chr12:92418926-92419817	MRPL42	CGTTCATGTTCACGATGTGG	CCACTTAACAGAGGGTCACCA	CATCTGGGTTACCATGTTGAAA	AAACTGAAGGGGAAGCAATG
Somatic deletion	chr2 :179023695 -1790241	PRKRA	CGAACTGAAAAGCAACACCA	GTCCTCCCCACAAAGGCTTA	GTATTGACTGCCAACCCACTC	TTAGGCCTCAACGACCCTAGAC

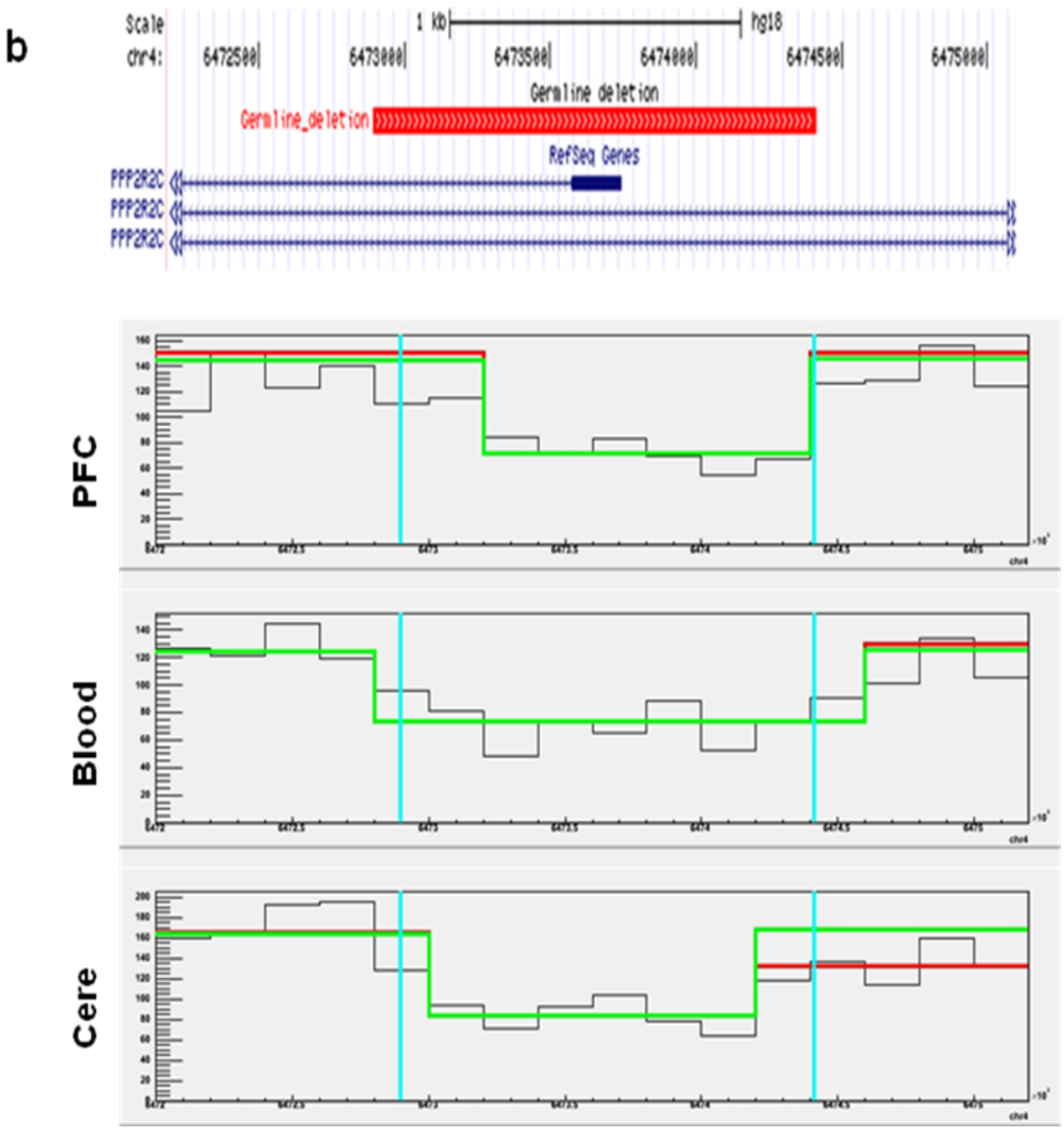
Somatic deletion	chr5:172967734 - 172969036	BOD1	CCCTGGGTTGC TGTAGTGTT	TGTGGGTGAGATTGT GCAGT	TAC TTG GTG TCA TGC CCT TG	GCT GCT GGT TGA GAA TTA CCC
Somatic deletion	chr7:26214778-26218067	CBX3	ATTCCCCCGGG TGTCTATTA	AACCAGTGCTATGGA TGCAA	GCTGGCAAAGAAAAAGAT GG	AAACCCCCAACAACTCT TCC
Somatic deletion	chr6:136641705-136642814	BCLAF1	CCCTCTACCCCT TCCTCTGT	GGGGCAGTCCGTAA AAATCT	CCCATAAGGTCGTCTCAT TCC	CCTGTCATGCAGGTGAA AAC
Somatic deletion	chr12:102902624-102903439	TDG	AGGCGGAGGCT CATTATTTT	CAATCCTGACCAAAC CGTCT	ACAAATTCAACCTTAAAA GCAACT	TACACATGTGGAGGGAA CCA
Somatic deletion	chr15:396522066-396522721	TYRO3	GAAGGAAAGGA AGGGGACAG	AGCCACTTGACAGG CAGTTT	GATATGGGAGCAGCCAG AGT	AGGCACAGCCTTGACG ATAG
Somatic deletion	chr3:67576394-67579498	SUCLG2	GTGGCCTTCAG CCTAATCAA	CTTTGAGTGCCTGGC ATTGT	ATGTGCATCCCCTTCACA AT	TGCCTAAAAAGACCTGC ACA
Somatic deletion	chr7:6986622-6992226	intergenic	CGCCAAGATGG GTAGATCAT	CCAACTCCAGTGTTC AAGCA	CCCATGGAGAAATCCCAT CT	TCCAGTGTTCAAGCAAT TTCC
Somatic deletion	chr12:102902624-102903439	TDG	ATATTCGCAGCC AGAGTGCT	AACAAACAGCAATGA TGCAGA	CTTCCCTCTACTCTGGCA CTTC	TCACTTTCCATGGCACA CTC



Supplementary Figure S1. Procedures for calling germline CNVs using sequencing data from three tissues from one individual. The number of candidates called at each step are in red. PFC, prefrontal cortex; B, blood; C, cerebellum.

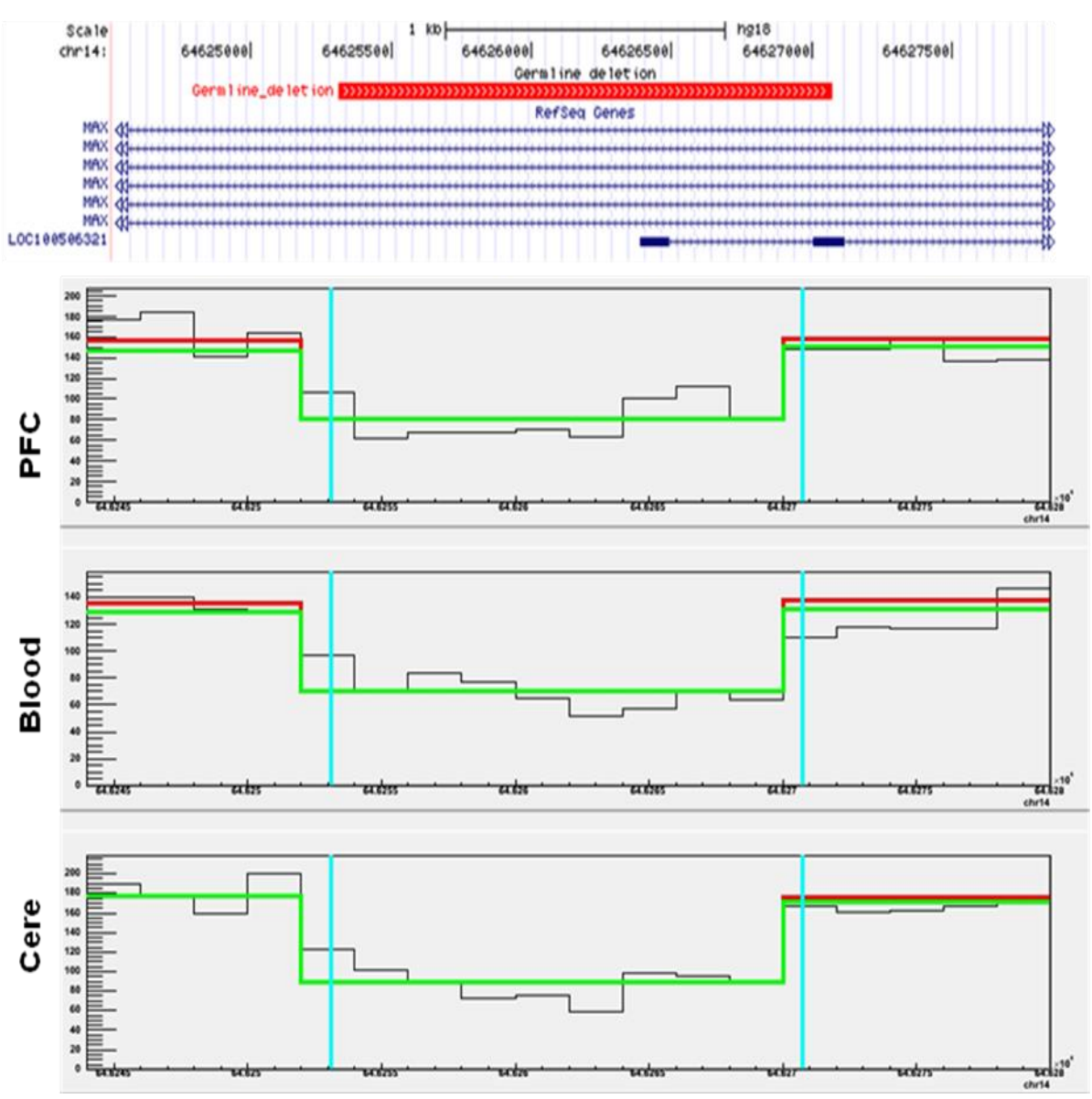

a
Scale
chr7:
500 bases
hg18
36408000
36408500
Germline deletion
Germline_deletion
RefSeq Genes
ANLN
PFC
Blood
Cere

Chr7 (*ANLN*)

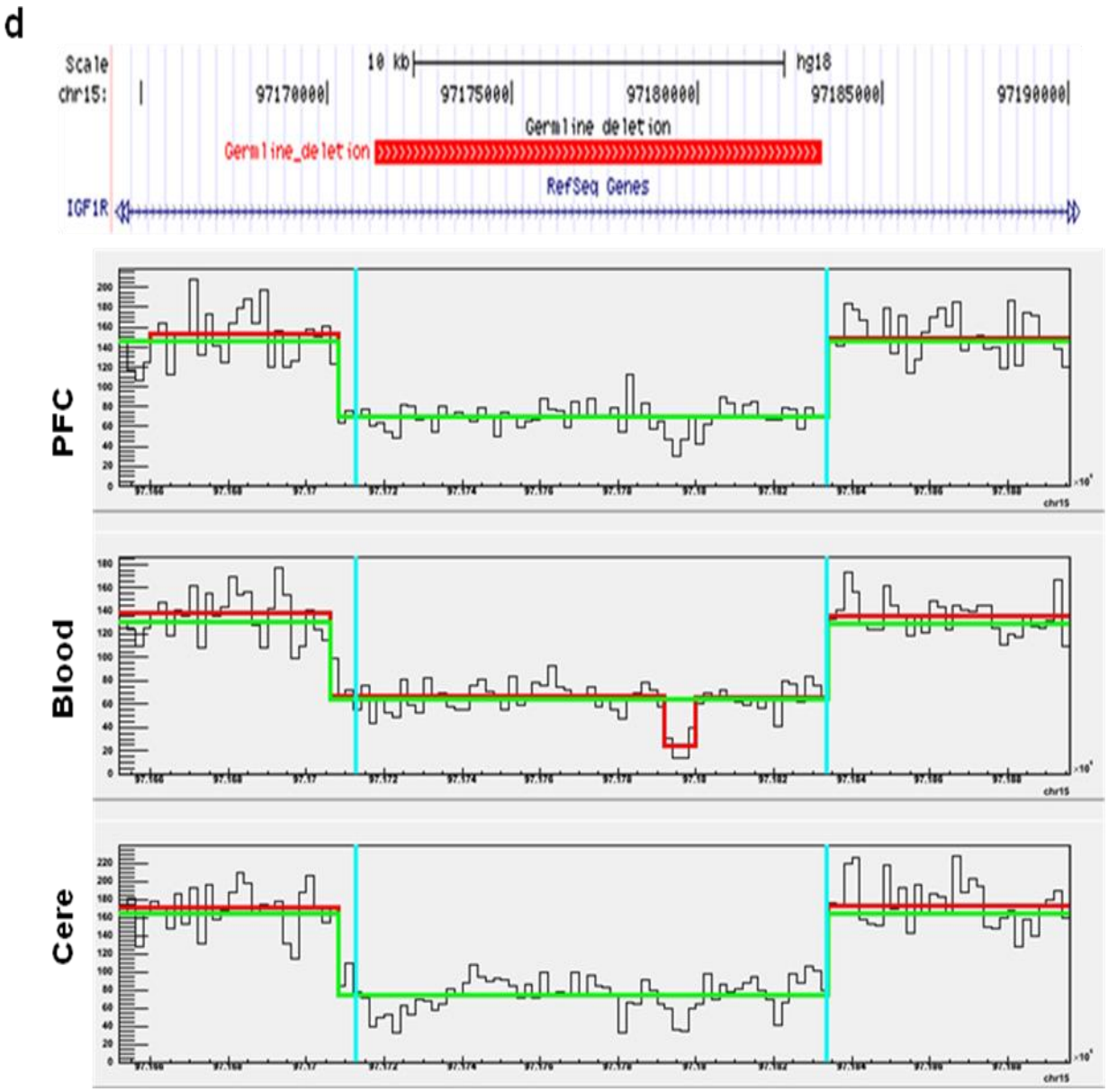


Chr4 (*PPP2R2C*)

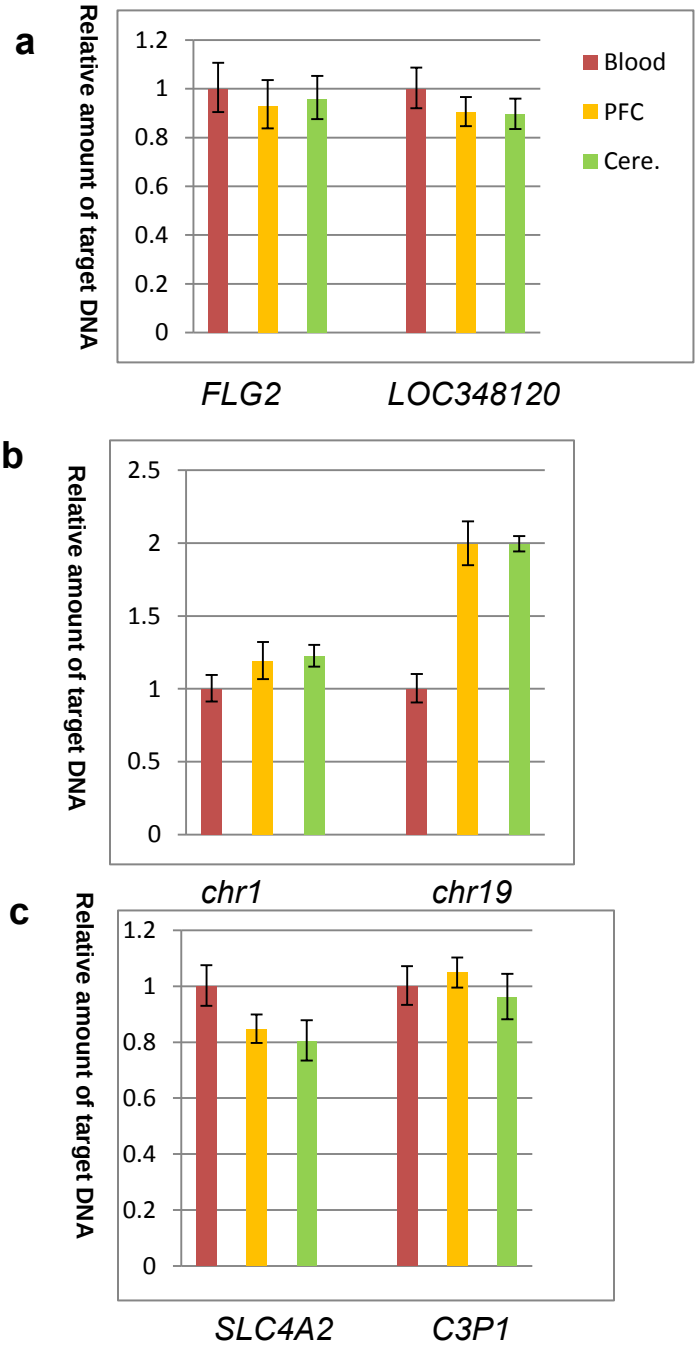
c


Scale
chr14:
1 kb
hg18
64625000
64625500
64626000
64626500
64627000
64627500
Germline deletion
Germline_deletion
RefSeq Genes
MAX
MAX
MAX
MAX
MAX
MAX
LOC100506321
PFC
Blood
Cere

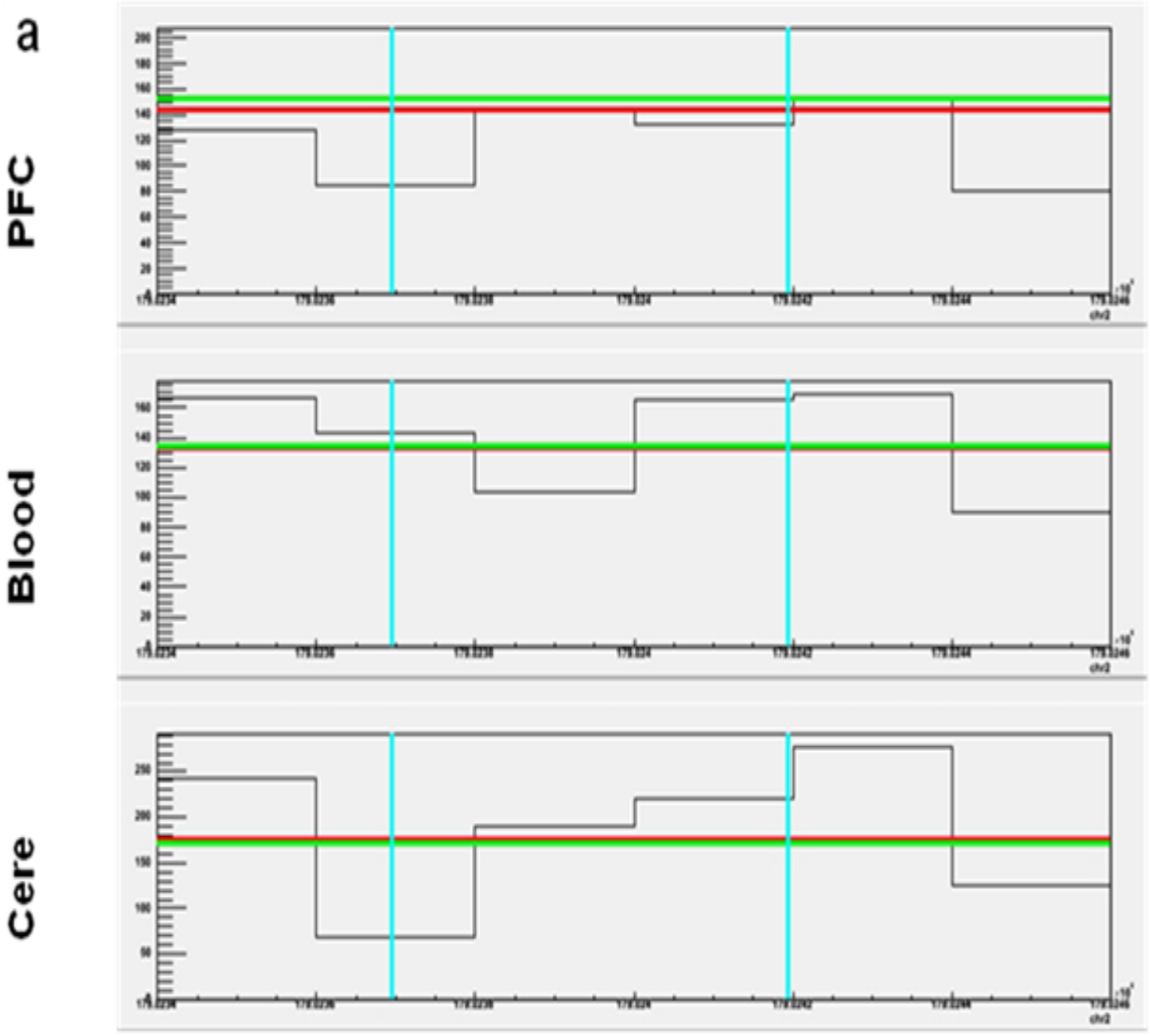
Chr14 (*MAX*)



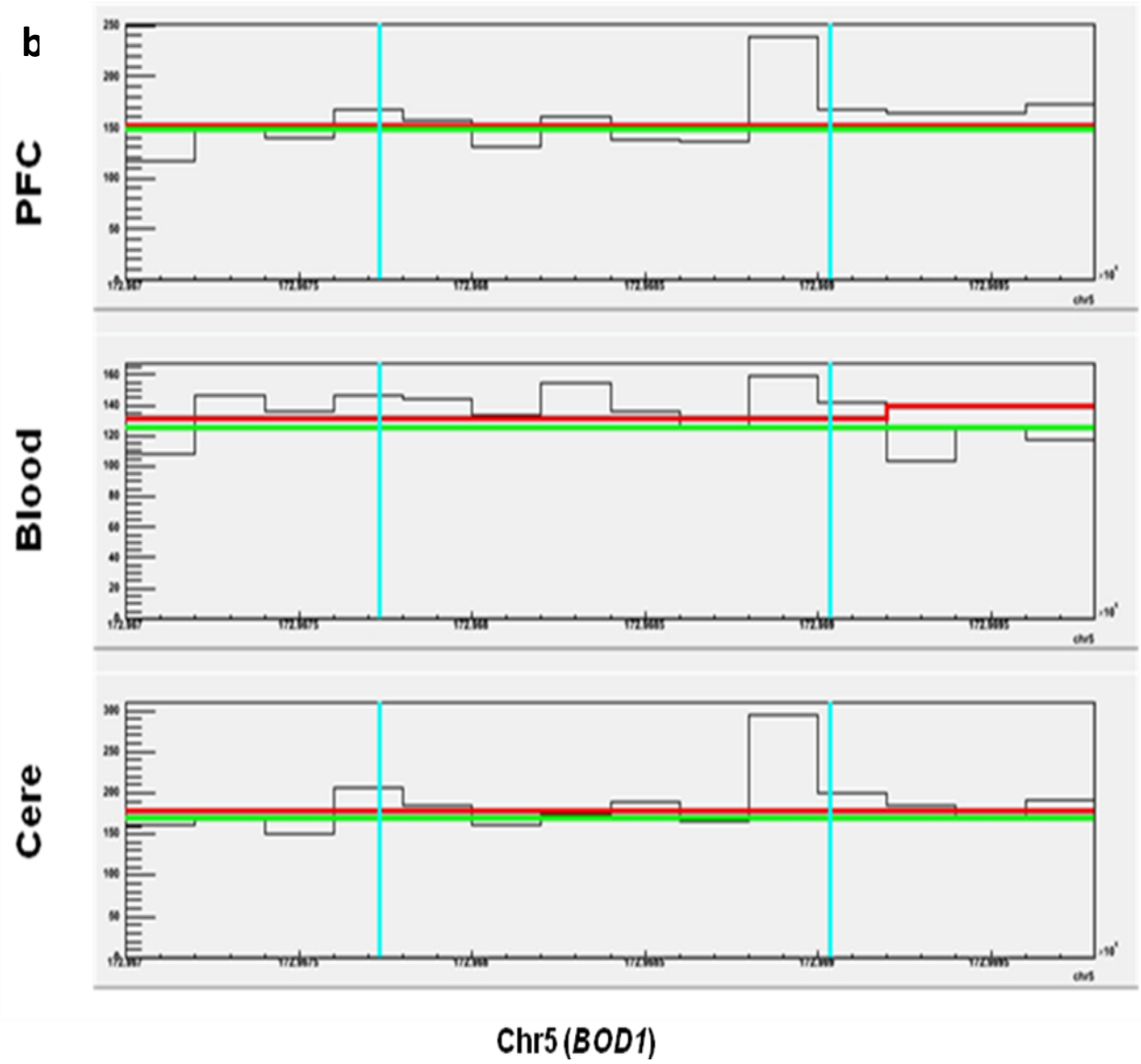
Supplementary Figure S2. Read depth of coverage of chromosome regions of 4 novel germline deletions. Read depth coverage shows a homozygous deletion in *ANLN* (a) and heterozygous deletions in *PPP2R2C*, *MAX* and *IGF1R* (b-d)



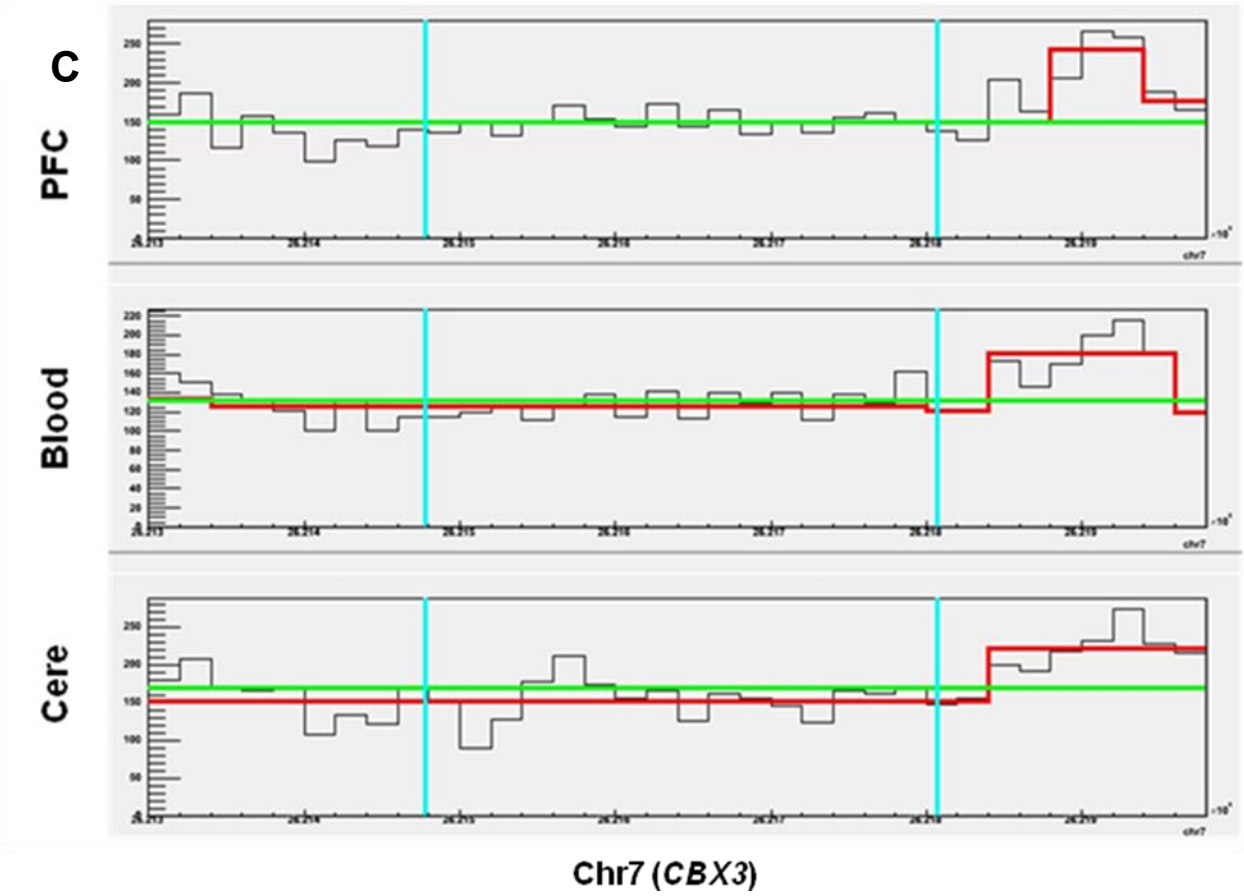
Supplementary Figure S3. qPCR of candidate CNV regions that were called using a read depth based mapping software, CNVnator. (a) two somatic duplication candidates specific to PFC, (b) two somatic duplication candidates specific to cerebellum, and (c) two somatic deletion candidates specific to cerebellum. The cerebellum specific deletion in *C3P1* was the only successful validation.



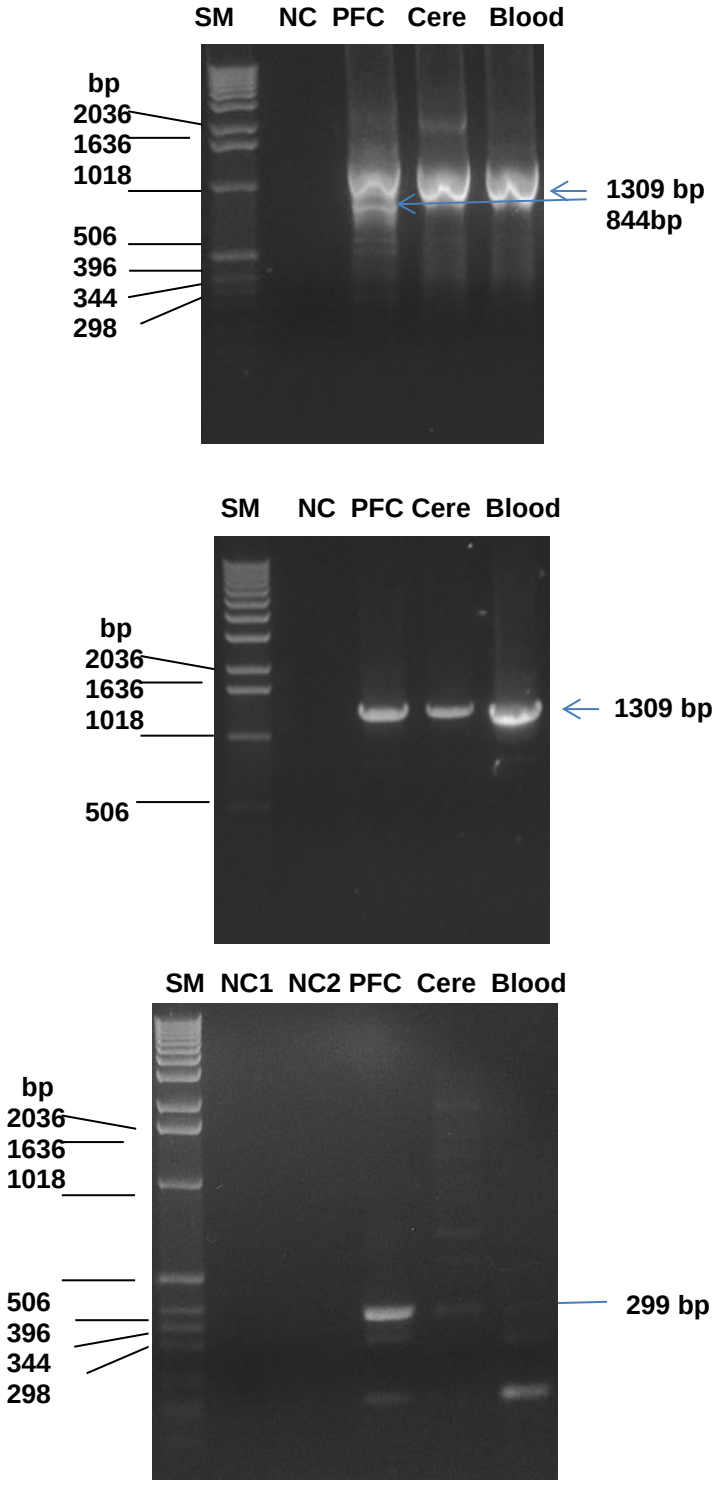
Chr2 (*MIR548N, PRKRA*)



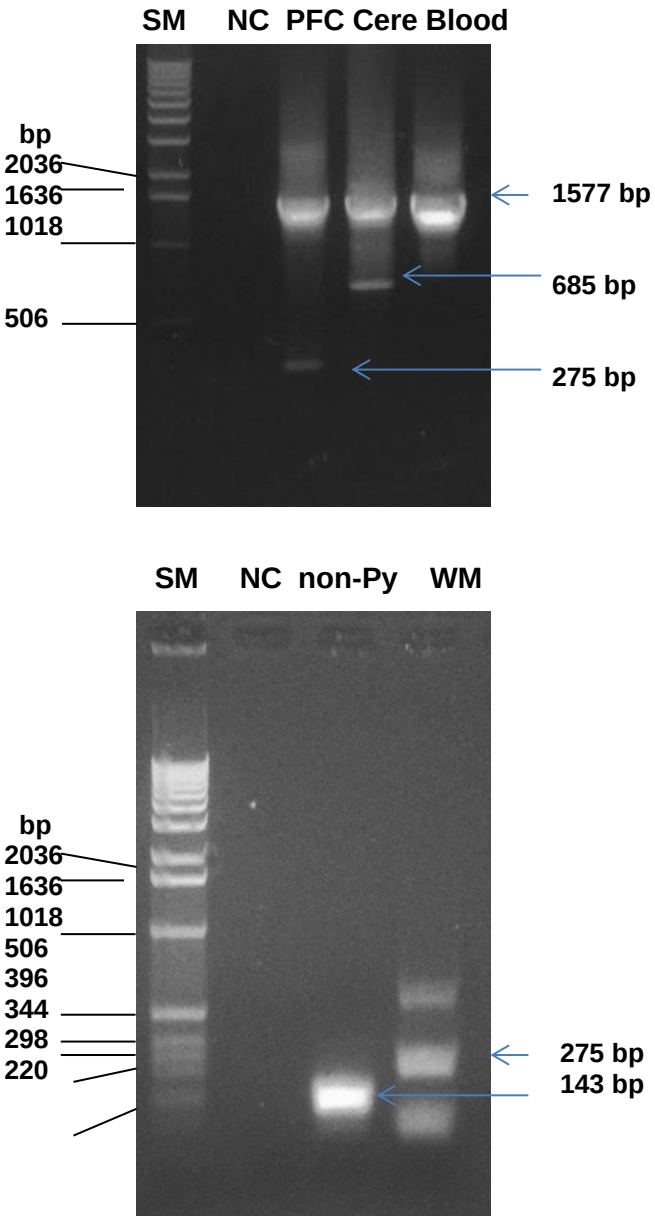
Chr5 (*BOD1*)



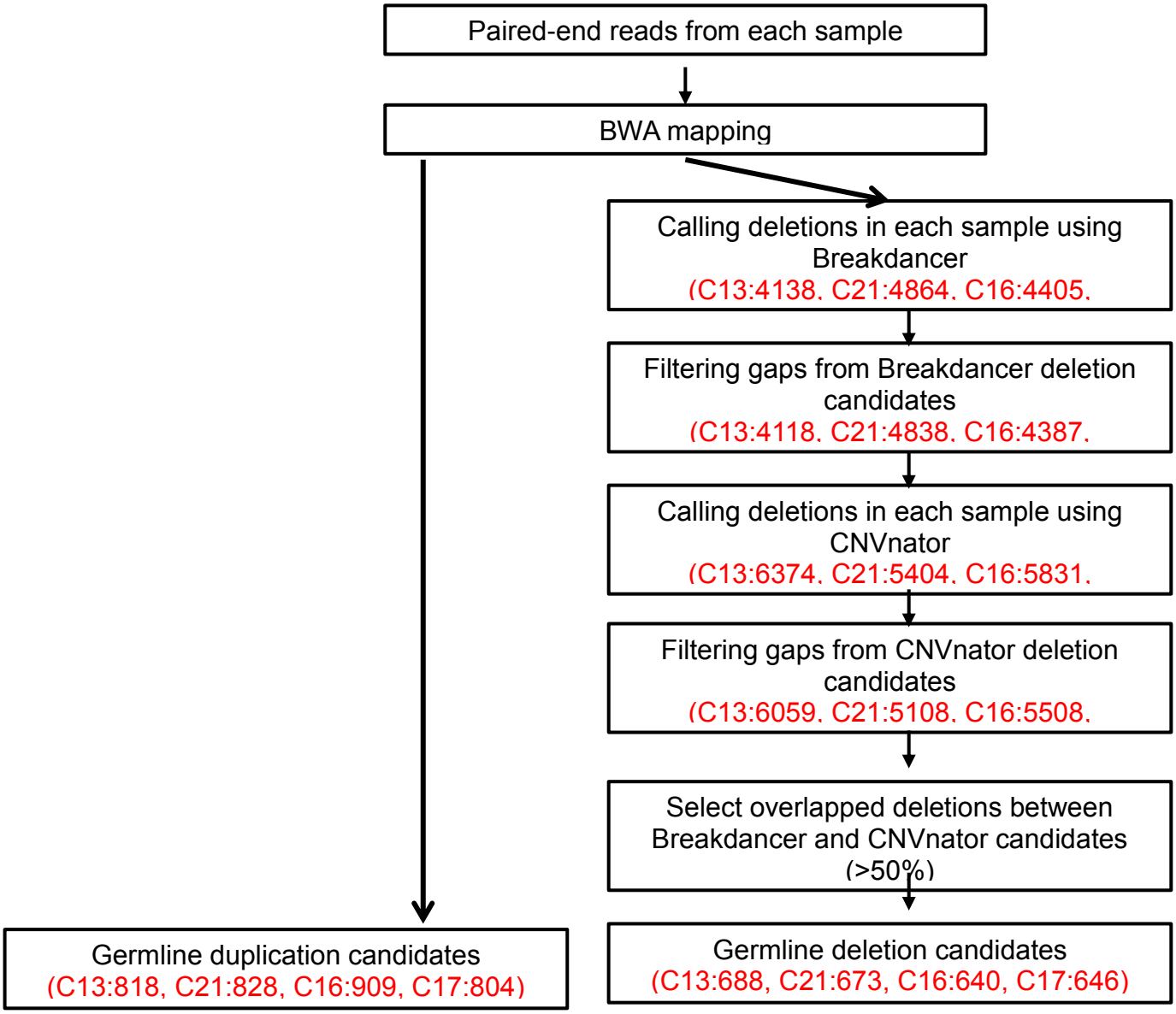
Supplementary Figure S4. Read depth of coverage of chromosome regions of somatic deletions in *PRKRA*, *BOD1* and *CBX3*. Unlike the germline deletions, the read depth analysis indicated that there were no clear declines in read depths in the somatic deletion regions.



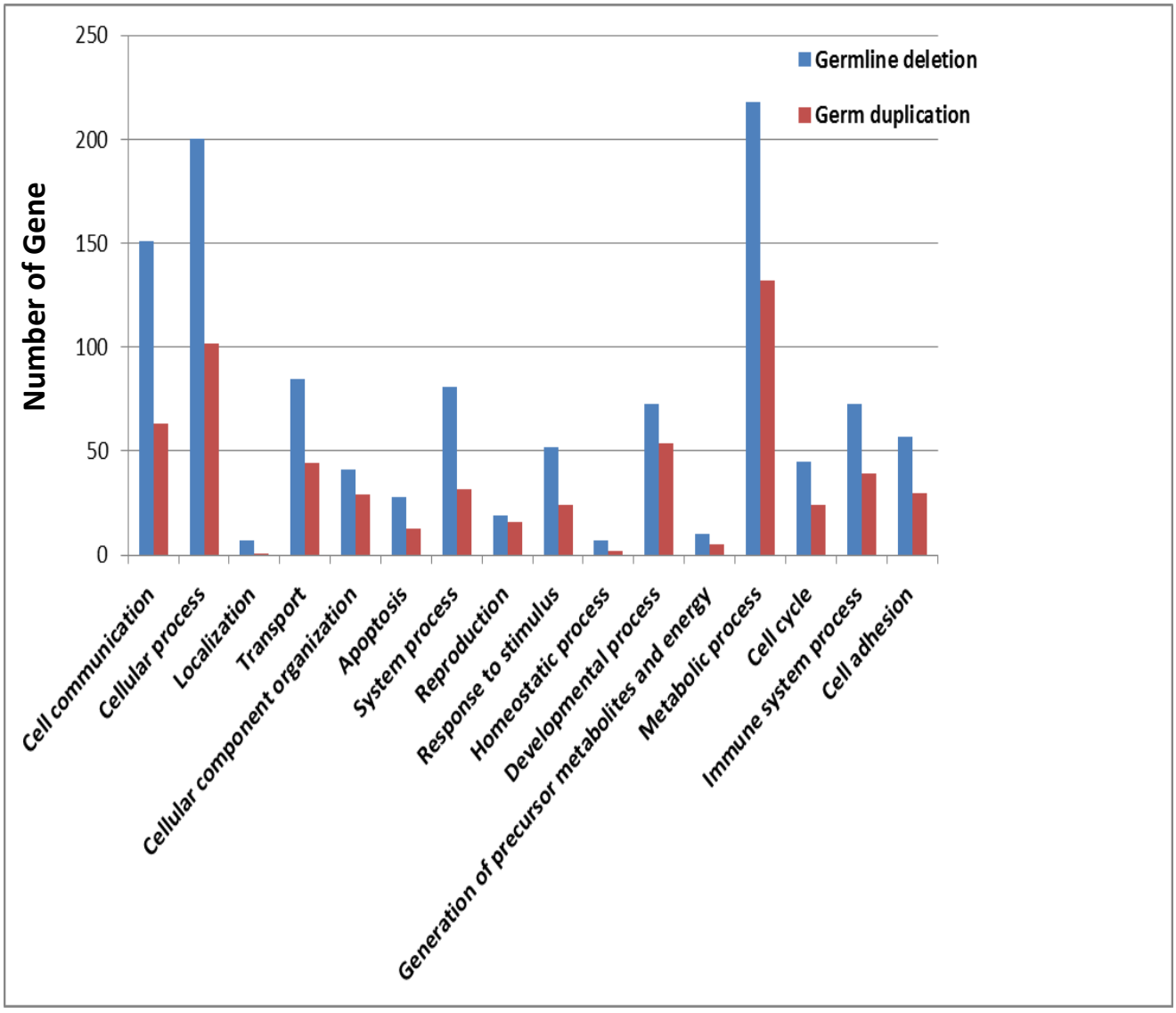
Supplementary Figure S5. Full images of agarose gels from Figure 2. NC1 molecular biology grade water was used for non-template negative PCR control. NC2, 1ul of non-template negative PCR product was used as a template for NC2. PFC: prefrontal cortex, Cere: cerebellum. DNA fragments which were sequenced are indicated by arrow.



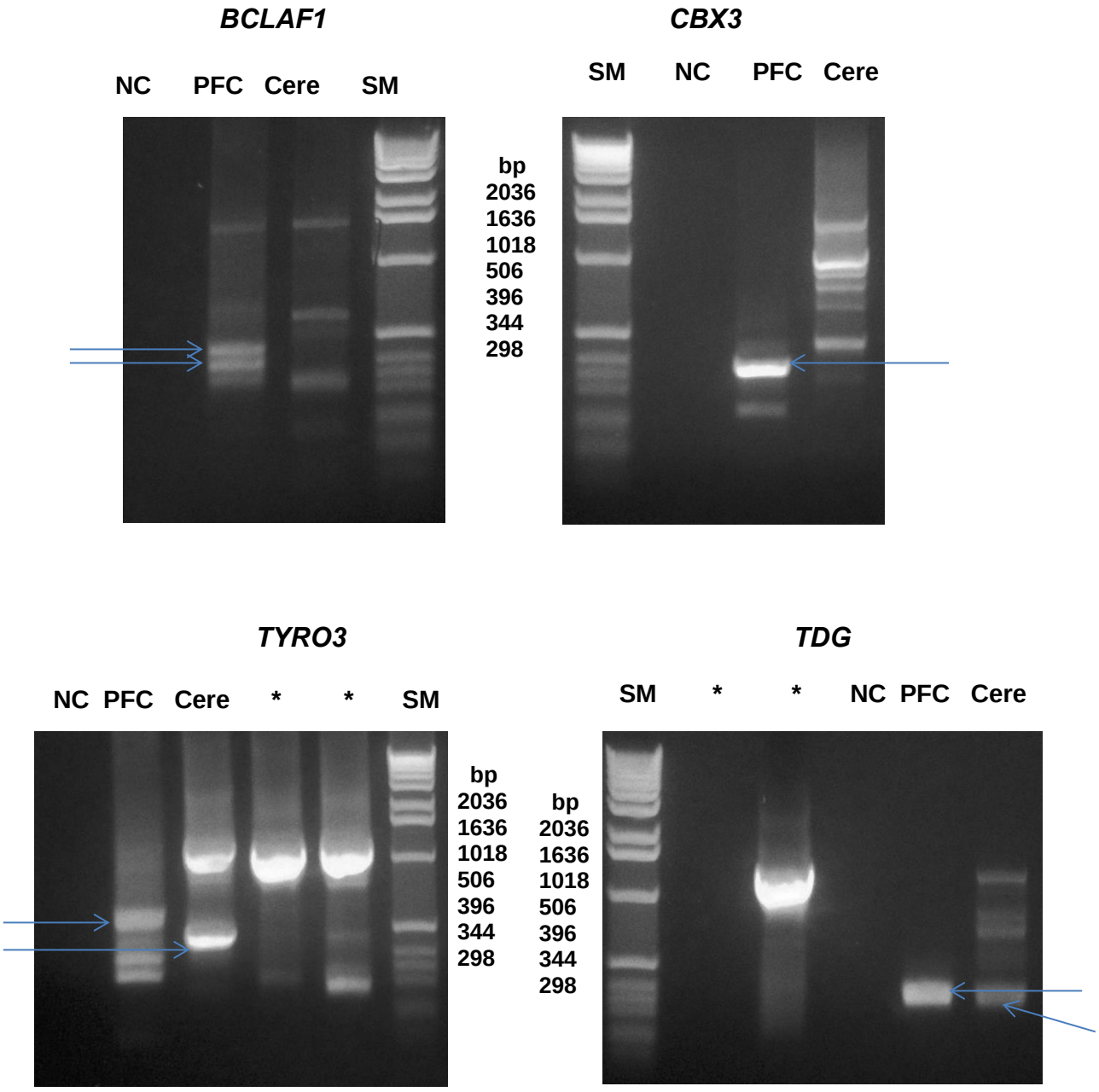
Supplementary Figure S6. Full images of agarose gels from Figure 3. NC: no template control, PFC: prefrontal cortex, Cere: cerebellum, non-Py; non-pyramidal cells, WM; cells in white matter. DNA fragments which were sequenced are indicated by arrow.

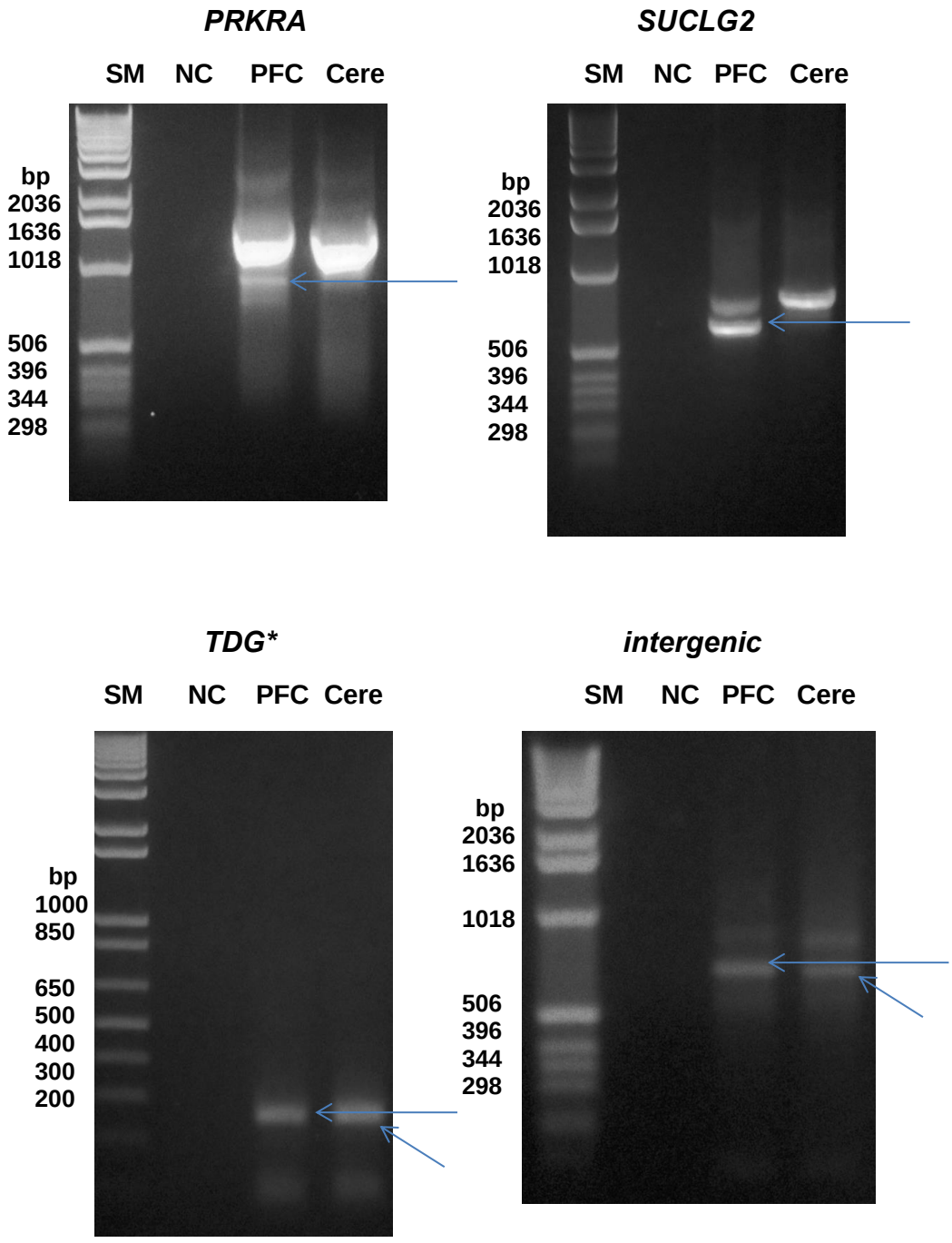


Supplementary Figure S7. Procedures for calling germline CNVs using sequencing data from single tissues from two individuals with schizophrenia and two unaffected controls. The number of candidates called at each step are in red. C13 and C21, unaffected controls; C16 and C17, schizophrenia

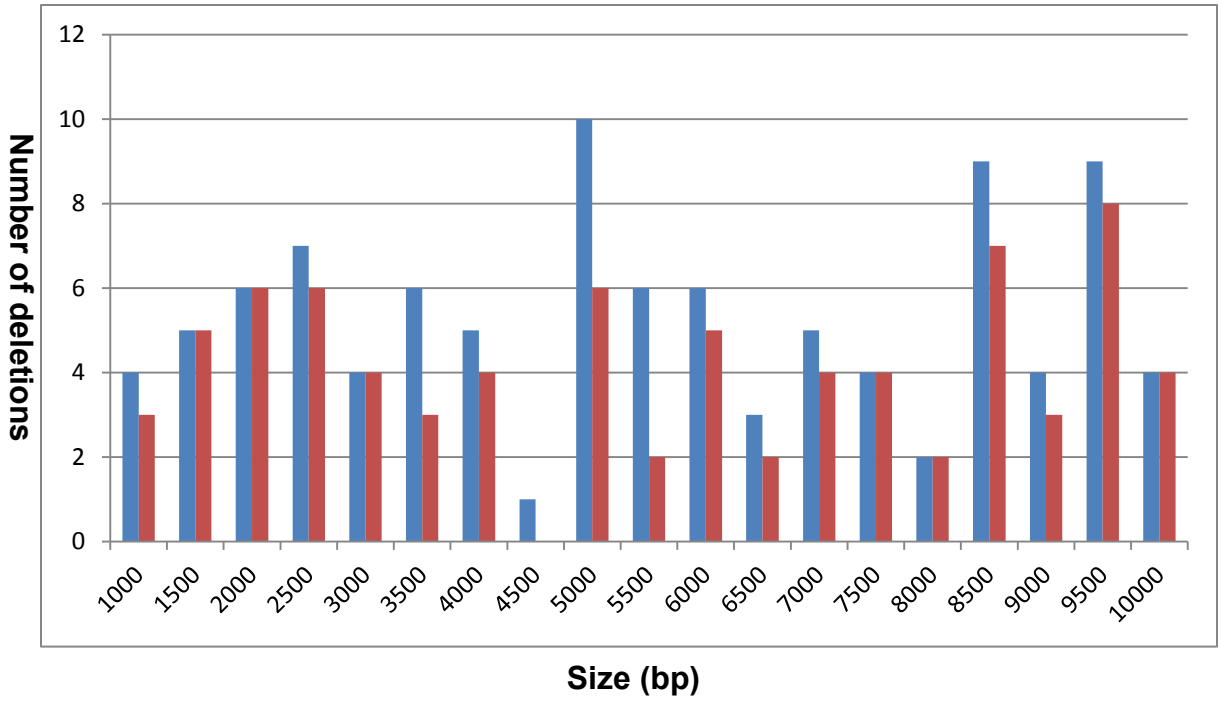


Supplementary Figure S8. Biological processes related to genes disrupted by germline CNVs in the PFC from two individuals with schizophrenia and two unaffected controls. Classification of the Gene Ontology biological processes was done by using Panther software.

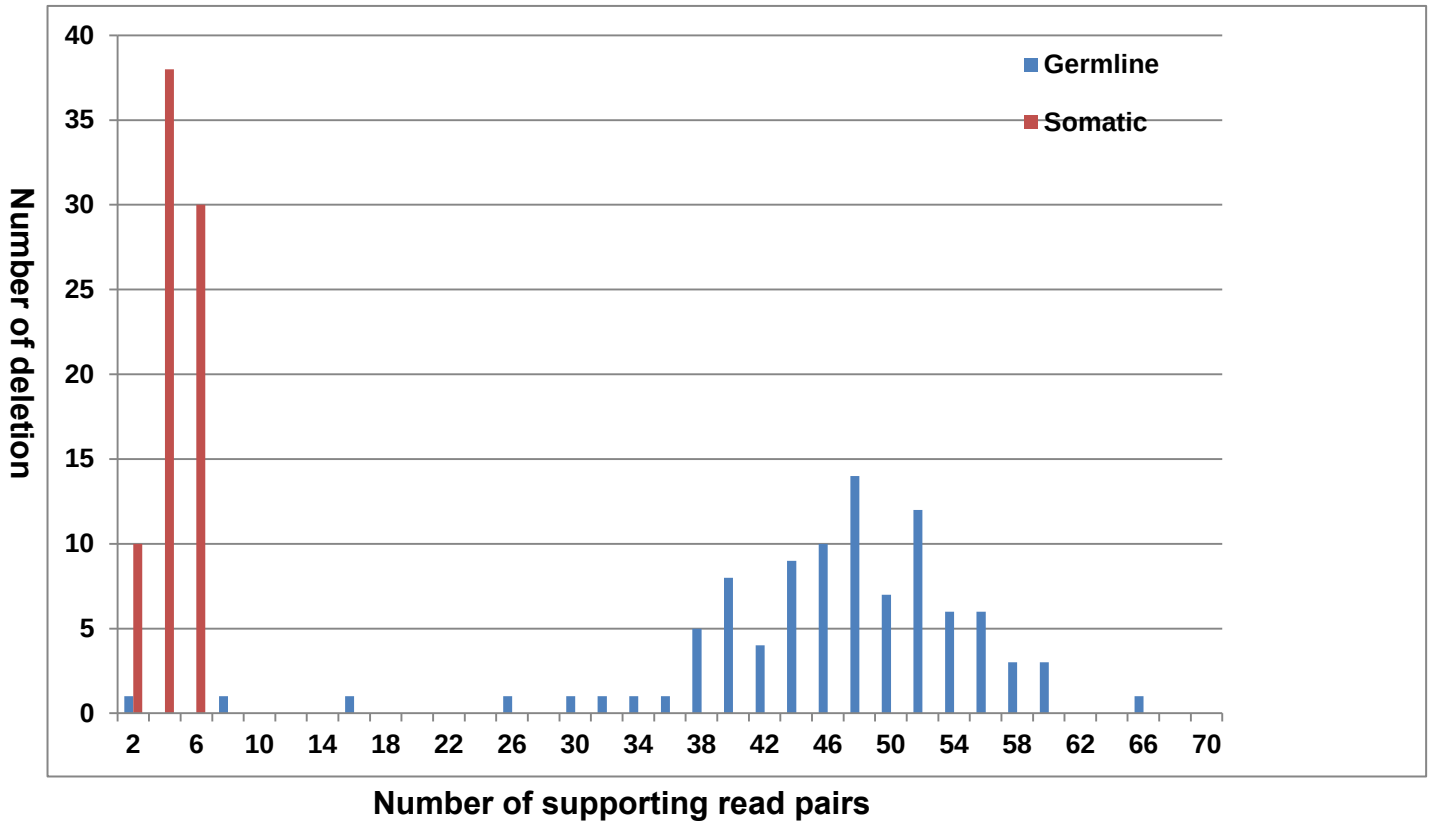

BCLAF1
NC
PFC
Cere
SM
bp
2036
1636
1018
506
396
344
298
CBX3
SM
NC
PFC
Cere
TYRO3
NC
PFC
Cere
*
*
SM
bp
2036
1636
1018
506
396
344
298
TDG
SM
*
*
NC
PFC
Cere
bp
2036
1636
1018
506
396
344
298



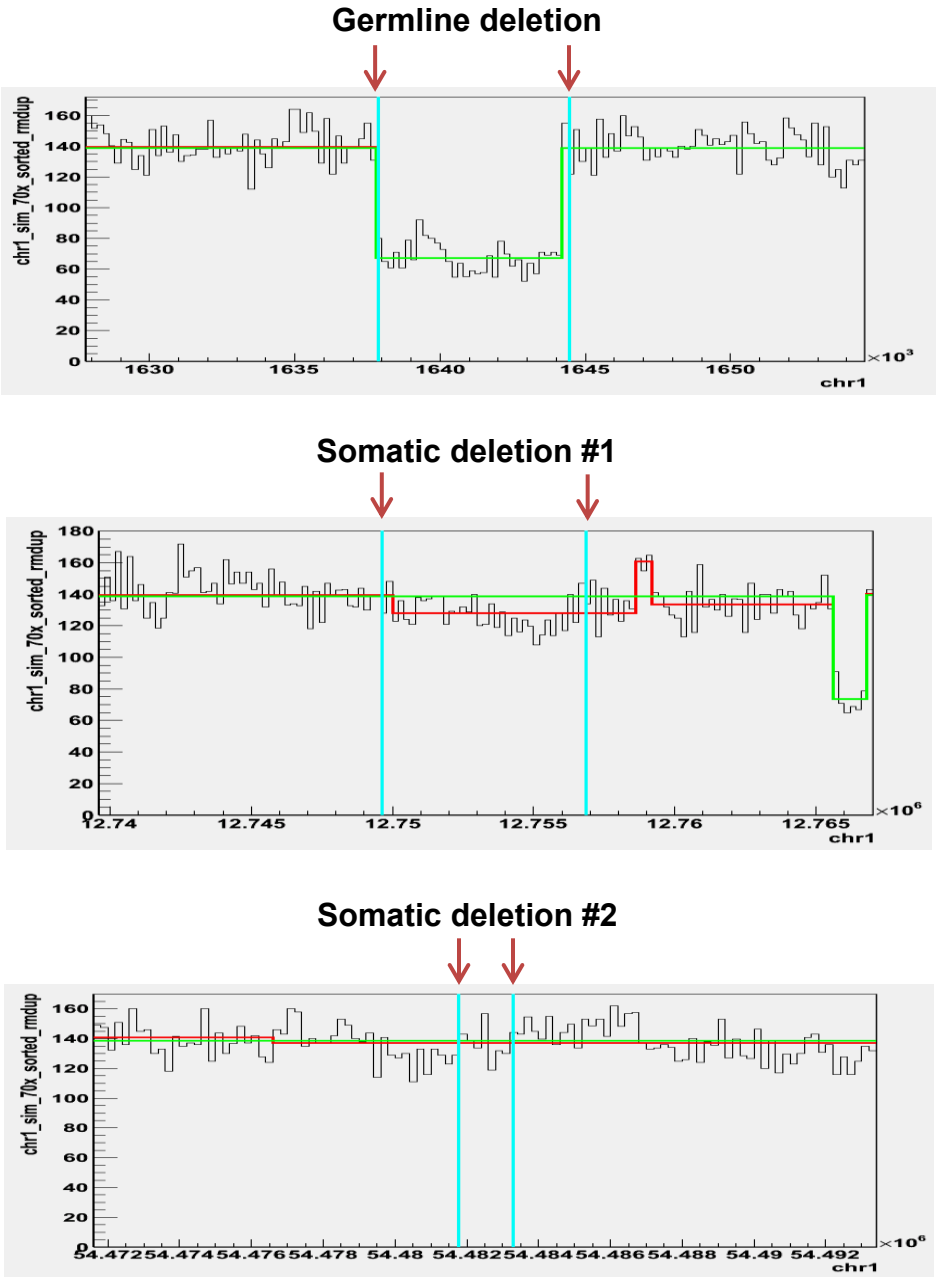
Supplementary Figure S9. Full images of agarose gels from Figure 5. NC: no template control, PFC: prefrontal cortex, Cere: cerebellum. DNA fragments which were sequenced are indicated by arrow. * DNA samples are independent to this experiment



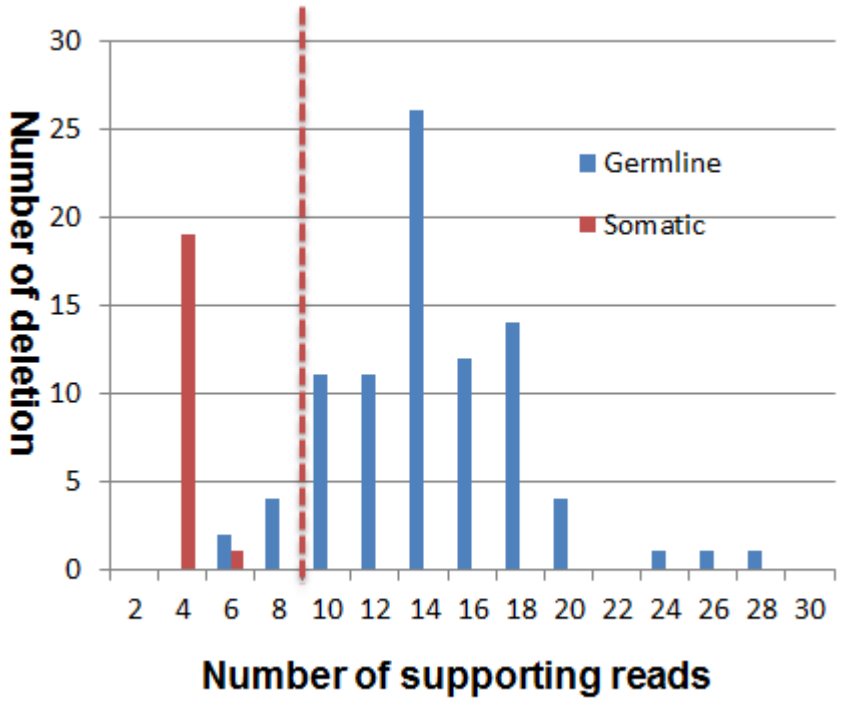
Supplementary Figure S10. The number of somatic deletions (total 100) that we generated in the simulated whole genome sequencing data (blue) compared to the number of somatic deletions (total 78) that we detected (red) in the simulated data using our integrated pipeline. There were no false positives detected.



Supplementary Figure S11. Number of supporting read pairs of detected germline deletions and somatic deletions using Breakdancer in the simulation experiment.



Supplementary Figure S12. Read depth coverage of chromosome regions of germline deletion with 2 supporting read pairs (top) and somatic deletions with 2 supporting read pairs (middle and bottom) in simulation experiment. Read arrows: breakpoints.



Supplementary Figure S13. Number of supporting reads of detected germline deletions and somatic deletions using Pindel in simulation experiment.

Supplementary Methods

Whole genome sequencing

For whole-genome sequencing (WGS), at least two DNA libraries were constructed to minimize the short-read redundancy of PCR duplicates, which could bias the read depth of sequencing coverage. DNA library preparation was carried out using Illumina, Inc., paired-end protocols. In brief, 1µg of genomic DNA isolation was fragmented to insert size between 200 to 700 bp using Covaris Acoustic Solubilizer (Covaris Inc.) with 20% duty cycle, 4 intensity, and 200 cycles per burst for 160 sec, at 4°C. These fragments were electrophoresed on a 2% agarose gel from where the 350–450 bp fraction was excised and extracted using the Qiagen gel extraction kit. The size-fractionated DNA was end repaired using T4 DNA polymerase, Klenow polymerase and T4 polynucleotide kinase. The resulting blunt-ended fragments were A-tailed using a 3′–5′ exonuclease-deficient Klenow fragment and ligated to Illumina paired-end adaptor oligonucleotides in a 'TA' ligation at room temperature for 15 min. The ligation mixture was electrophoresed on a 2% agarose gel and size-selected by removing a 2-mm horizontal slice of gel at ~500 bp using a sterile scalpel blade. DNA was extracted from the agarose as above. 10ng of the resulting DNA was PCR-amplified for 18 cycles using 2 units of Phusion polymerase. PCR cleanup was performed using AMPure beads (Agencourt BioSciences Corporation) following the manufacturer's protocol.

Paired-end read alignment

Paired-end reads were mapped to the hg18 human reference genome using Burrows-Wheeler Aligner (BWA)[1] version 0.5.9rc1. Mapped SAM files were merged into one main BAM file for each sample, PFC, cerebellum and blood, and for the 4 additional PFC samples. Sorting and removing potential duplicates was performed by SAMtools software package[2] (0.1.13). Three resulting BAM files from PFC, cerebellum, and blood were used for the discovery phase analysis, and the four BAM files from the additional PFC samples were used for exploratory phase analysis. Basic statistics for mapped data quantities were obtained by SAMtools and GATK[3].

Identifying germline deletion and duplication

In the first discovery phase, BreakDancer[4] and CNVnator[5] were used to identify germline deletions in our data. Only CNVnator[5] was used to call germline duplications. Minimum mapping quality of 35 and cutoff standard deviation of four were used to detect structural variations from BreakDancer. Candidate regions were determined to be germline if they met the following two criteria. First, if the candidate regions in PFC and cerebellum had a called deletion that overlapped more than 50% with one another. Second, the PFC candidate region also had to overlap more than 50% of the region called in the blood. CNVs in blood tissue were used as a baseline reference variation to call germline CNVs. To remove false positives due to the performance of each variant calling tool, we extracted the

results from both tools that overlapped. The methodology for determining germline deletions was identical to that used to determine if candidates were germline duplications. Thus, a germline duplication was called if both duplications from prefrontal cortex and cerebellum overlapped more than 50% and if the PFC duplication also overlapped more than 50% of the duplication called from blood DNA.

In the second phase, we called germline CNVs in data from a single tissue without comparing data from multiple tissues. Basically if a CNV is present in more than 50% of the sequenced genomes, we assumed that the CNV is a germline variation. We determined germline deletion candidates as those called with both tools, BreakDancer and CNVnator and the deletion sequence overlapped at least 50%. As BreakDancer does not support a duplication calling function, only CNVnator was used to identify germline duplications.

In the exploratory phase, we reported germline CNV candidates from CNVnator duplication and deletion calls for four additional PFC samples without additional processing.

Identifying brain-specific somatic deletion candidates in the first discovery phase

We set a conservative computational analysis pipeline to identify brain specific somatic deletions. A brain specific somatic deletion was first called, using Breakdancer, if the deletion in one brain area did not overlap with a deletion in the

other area or in blood at all. Then if, for example, a PFC specific somatic deletion was called and there were no overlapping deletions in cerebellum or blood at all, we would then filter out candidates that were initially called as deletions at the same genomic locus in the other tissues using Pindel [6]. We further filtered all false positive candidates that showed a clear decline in read depth in the other tissues by visual inspection using CNVnator viewer. For calling PFC and cerebellum common somatic deletions, we identified deletion candidates if both deletions in PFC and cerebellum overlapped more than 50% mutually and did not overlap more than 50% any deletions from the blood. We then filtered false positive candidates if a deletion overlap was called in blood using Pindel or a candidate deletion showed a clear decline in read depth in blood by using CNVnator viewer. We called 60 somatic deletions specific to PFC, 34 specific to cerebellum and 41 common to both PFC and cerebellum but absent in blood using the pipeline. We validated the brain specific somatic deletions in 3 genes. However, we were unable to validate 13 somatic deletion candidates. Among the false positives, 8 were germline deletions which were present in all tissues we tested and were detected in our initial PCRs. Therefore, to determine why the false discovery rate (FDR) was so high we conducted further analysis by performing a manual inspection of the chromosomal regions and read depth of the candidates to find the potential cause of these false positives. We found repetitive DNA sequences highly homologous to 500bp (1X library insert size) upstream of the left breakpoint and/or 500bp downstream of the right breakpoint of the deletion and were within 1X deletion size

from deletion breakpoint of four of the false positive candidates. We developed a pipeline to filter out such repetitive sequences which are likely to contribute to false deletion calls in paired end mapping as well as background noise in read depth based mapping [7]. The size of 6 of the 8 candidates which were actually germline deletions rather than somatic as originally called were relatively small (<400 bp). Thus, such small size candidates were not called as germline deletions in our initial analysis using the read depth based mapping software, CNVnator, and consequently this is a less reliable method for detecting small CNVs than paired end mapping [8]. We therefore filtered out relatively small size candidates (<400bp) in our analysis pipeline to reduce the FDR. Furthermore, visual inspection was also performed to remove false positive candidates that showed a clear deletion pattern in read depth by using CNVnator with 200-bp windows but not called deletions by the software. The Blat filter and size filter were added to our somatic deletion calling pipeline for sequencing data from multiple tissues from the same individual to reduce false positive findings**.**

Identifying somatic deletion candidates in the second phase

Validating candidates identified in the first discovery phase would guarantee the presence of brain-specific somatic deletions. Thus, in the second phase, we searched for somatic deletion candidates in data obtained from the PFC of two schizophrenic cases and two unaffected control samples. If a germline deletion occurs, all the cells in the human body have the variation. For a homozygote

germline deletion both alleles will have the variation whereas for a heretozygote germline deletion only one allele will have the variation. In contrast, somatic deletions which can occur in any stage of brain development, will only affect a fraction of brain cells. Thus, only a fraction of brain cells will have the somatic deletion and the deletion will be present in less genomic DNA extracted from the brain tissue than a heterozygote germline deletion that will be present in half of the genomic DNA. Consequently a somatic variation can be detected in WGS data from a single brain tissue by modifying the parameters of the individual programs used in our calling pipeline that we used in the first discovery phase of the study. Theoretically, somatic deletions that occur in a small fraction of cells may be called by less supporting read pairs in paired end mapping software, Breakdancer, and by less supporting reads in split read mapping software, Pindel, than germline deletions. The supporting read pairs are pairs with anomalous spacing in paired end mapping. The supporting reads in split mapping are reads spanning the breakpoints. While germline deletions can be called in read depth analysis using CNVnator, the somatic deletions cannot be called in the analysis. Thus in our pipeline, BreakDancer was first used to identify somatic deletion candidates just as we did in the first discovery phase. However, while we considered all candidates called by BreakDancer in the first phase, we only considered somatic deletion candidates if they were called by less than 7 anomalous spacing read pairs. We determined this threshold based on the results from validation experiments in the first phase. Then, to remove potential germline deletions from the somatic deletion

candidates called using Breakdancer, we filtered out those deletion candidates that had at least 50% mutual overlap with deletions called from Pindel and that had more than 9 supporting reads. We determined this threshold also based on the results from validation experiment in the first phase. To further filter out potential germline deletions, we then used germline calls from CNVnator. To remove false positive candidates that occurred by misguided mapping of paired reads, we performed the Blat filtering process. Visual inspection was also performed to remove false positive candidates that showed a clear deletion pattern in read depth by using CNVnator with 200-bp windows but not called by the software.

Blat filtering

According to the characteristics of paired-end sequence data, misguided multiple mapping of paired reads occurring by sequence homology causes false positive variation detection. Previous approaches have used strict mapping quality or supporting read numbers to overcome these errors [7]. However, we predicted that somatic deletion candidate detection relied on detecting candidates with only a few supporting reads. To remove false positives without excluding such uncertain candidates, we utilized Blat alignment tool to find false positives that occurred by sequence homology. For each deletion candidate, we read a 100-bp sequence (a single read size) near the breakpoint and searched for homologous sequences. We filtered out deletion candidates if homologous sequences larger than 90-bp were found located in one insert size from another breakpoint. Searching

homologous sequences was performed by standalone Blat v.34x13, and the overall filtering process was implemented by JAVA.

Validating breakpoints of germline and somatic deletions by PCR and Sanger sequencing

We used amplified chromosomal DNA for validation of germline deletions and for initial validation of somatic deletions as limited amounts of DNA were available from the same batch of extractions. For whole genome amplification, we used 50ng of input DNA using REPL-g whole genome amplification kit (Qiagen). Deleted DNA fragments from the 4 novel germline deletions were amplified by PCR using 25ng of DNA as a template. The specific primers for each deletion were designed using Primer3[9]. PCR primers are listed in **Supplementary Table S9.** PCR reaction mixtures contained 2.5Unit Taq DNA polymerase (Qiagen), 1X PCR buffer, 1X Q solution (Qiagen), 0.2mM dNTP and 0.4 µM of each primer. PCR conditions were as follows: 94^{o}C for 3 minutes, then 35 cycles of 94^{o}C for 30 seconds, 55^{o}C for 30 seconds, 72^{o}C for 2 minutes, followed by 72^{o}C for 7 minutes. PCR products were sequenced using the Sanger method. For validation of somatic deletions, first round PCRs were done using 100ng of DNA as template. PCR reactions were performed with or without 1X Q solution (Qiagen). PCR conditions were the same as the conditions for germline deletions described above. Nested PCR was then done using 1µl of first round PCR product as a template. PCR conditions were the same as described above. If a deleted DNA fragment was visualized on agarose

gel by nested PCR, the DNA fragments were sequenced after gel extraction (Qiagen). If a deleted DNA fragment was not visualized on agarose gel, nested PCR was re-performed using eluted DNA from first round PCR as a template. For extracting DNA fragments from agarose gel, first round PCR products were resolved by agarose gel electrophoresis. The agarose gel fragment was excised at the expected sizes of the deleted fragments and DNA was extracted using MinElute Gel Extraction kit (Qiagen). For a reconfirmation of PFC specific somatic deletion in *PRKRA* gene using unamplified chromosomal DNA, first round PCR was done using 10ng of DNA as a template with 1X Q solution. PCR conditions were as follows: 94^{o}C for 3 minutes, then 35 cycles of 94^{o}C for 30 seconds, 60^{o}C for 30 seconds, 72^{o}C for 1 minute, followed by 72^{o}C for 10 minutes. PCR primers were BPF (5'-CCCTTCCCGGAGCTACGGC-3') and Primer R (5'-GTCCTCCCCACAAAGGCTTA-3'). Nested PCR was done using 1µl of the first round PCR product as a template. PCR conditions were as follows: 94^{o}C for 3 minutes, then 35 cycles of 94^{o}C for 30 seconds, 55^{o}C for 30 seconds, 72^{o}C for 1 minute, followed by 72^{o}C for 10 minutes. Primers for nested PCR were BP nested F (5'-GCTCCGCCCCCACCCTGC-3') and nested R (5'-TTAGGCCTCAACGACCCTAGAC-3'). PCR products were sequenced using the Sanger method. If multiple PCR products of different sizes are generated, DNA fragments of expected size were mainly sequenced.

References

1 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
2 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
3 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
4 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-681 (2009).
5 Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**, 974-984 (2011).
6 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).
7 Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36-46 (2012).
8 Shen, Y., Gu, Y. & Pe'er, I. A hidden Markov model for copy number variant prediction from whole genome resequencing data. *BMC Bioinformatics* **12 Suppl 6**, S4 (2011).
9 Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).